SMCHD1 mutations associated with a rare muscular dystrophy can also cause isolated arhinia and Bosma arhinia microphthalmia syndrome

Mutations in *SMCHD1* are Associated with Isolated Arhinia, Bosma Arhinia Microphthalmia Syndrome, and Facioscapulohumeral Muscular Dystrophy Type 2

Natalie D Shaw[1,2*], Harrison Brand[1,3-5*], Zachary A Kupchinsky[6], Hemant Bengani[7], Lacey Plummer[1], Takako I Jones[8], Serkan Erdin[3,5], Kathleen A Williamson[7], Joe Rainger[7], Kaitlin Samocha[5,9], Alexei Stortchevoi[3], Benjamin B Currall[3], Ryan L. Collins[3,10], Jason R Willer[6], Angela Lek[11], Monkol Lek[5,9], Malik Nassan[12], Shahrin Pereira[13], Tammy Kammin[13], Diane Lucente[3], Alexandra Silva[3], Catarina M Seabra[3,14], Yu An[3], Morad Ansari[7], Jacqueline K Rainger[7], Shelagh Joss[15], Jill Clayton Smith[16], Margaret F Lippincott[1], Sylvia S. Singh[1], Nirav Patel[1], Jenny W Jing[1], Jennifer Law[17], Nalton Ferraro[18], Alain Verloes[19], Anita Rauch[20], Katharina Steindl[20], Markus Zweier[20], Ianina Scheer[21], Daisuke Sato[22], Nobuhiko Okamoto[23], Christina Jacobsen[24], Jeanie Tryggestad[25], Steven Chernausek[25], Lisa A Schimmenti[26], Benjamin Brasseur[27], Claudia Cesaretti[28], Jose E. García-Ortiz[29], Tatiana Pineda Buitrago[30], Orlando Perez Silva[31], Jodi D Hoffman[32], Wolfgang Mühlbauer[33], Klaus W Ruprecht[34], Bart Loeys[35], Masato Shino[36], Angela Kaindl[37], Chie-Hee Cho[38], Cynthia C Morton[5,13], Veronica van Heyningen[7], Eric C Liao[39], Ravikumar Balasubramanian[1], Janet E Hall[1,2], Stephanie B Seminara[1], Daniel Macarthur[5,9,40], Steven A Moore[41], Koh-ichiro Yoshiura[42], James F Gusella[3-5,11], Joseph A Marsh[7], John M Graham, Jr[43], Angela E Lin[43], Nicholas Katsanis[6], Peter L Jones[8], William F Crowley, Jr[1], Erica E Davis[6**], David R FitzPatrick[7**], Michael E Talkowski[3-5,40**]

[1.] Harvard Reproductive Endocrine Sciences Center and NICHD Center of Excellence in Translational Research in Fertility and Infertility, Reproductive Endocrine Unit of the Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA
[2.] National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA
[3.] Molecular Neurogenetics Unit and Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA
[4.] Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA
[5.] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
[6.] Center for Human Disease Modeling, Duke University Medical Center, Durham, North Carolina, USA
[7.] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh Western General Hospital, Edinburgh, UK
[8.] Department of Cell and Developmental Biology, University of Massachusetts Medical School, Worcester, Massachusetts, USA
[9.] Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA
[10.] Program in Bioinformatics and Integrative Genomics, Division of Medical Sciences, Harvard Medical School, Boston, Massachusetts, USA
[11.] Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA
[12.] Department of Psychiatry and Psychology, Mayo Clinic, Rochester, Minnesota, USA
[13.] Department of Obstetrics, Gynecology, and Reproductive Biology, Brigham and Women's Hospital, Boston, Massachusetts, USA
[14.] GABBA Program, University of Porto, Porto, Portugal
[15.] West of Scotland Genetics Service, South Glasgow University Hospitals, Glasgow, UK
[16.] Faculty of Medical and Human Sciences, Institute of Human Development, Manchester Centre for Genomic Medicine, University of Manchester, Manchester Academic Health Science Centre (MAHSC), Manchester, UK
[17.] Division of Pediatric Endocrinology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[18.] Department of Oral and Maxillofacial Surgery, Boston Children's Hospital, Boston, Massachusetts, USA
[19.] Department of Genetics, Robert Debré Hospital, Paris, France

[20.] Institute of Medical Genetics, University of Zurich, Schlieren-Zurich, Switzerland.
[21.] Department of Diagnostic Imaging, Children's Hospital, Zurich, Switzerland
[22.] Department of Pediatrics, Hokkaido University Graduate School of Medicine, Sapporo, Japan
[23.] Department of Medical Genetics, Osaka Medical Center and Research Institute for Maternal and Child Health, Osaka, Japan
[24.] Division of Endocrinology and Genetics, Children's Hospital Boston, Boston, Massachusetts, USA
[25.] Department of Pediatrics, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, USA
[26.] Departments of Otorhinolaryngology and Clinical Genomics, Mayo Clinic, Rochester, Minnesota, USA
[27.] DeWitt Daughtry Family Department of Surgery, University of Miami Leonard M Miller School of Medicine, Miami, Florida, USA
[28.] Medical Genetics Unit, Fondazione IRCCS Ca` Granda, Ospedale Maggiore Policlinico, Milan, Italy
[29.] División de Genética, Centro de Investigación Biomédica de Occidente-Instituto Mexicano del Seguro Social, Guadalajara, Jalisco, México
[30.] Fundación Hospital Infantil Universitario de San José, Bogota, Columbia
[31.] Private Plastic Surgery practice, Bogota, Columbia
[32.] Division of Genetics and Division of Maternal Fetal Medicine, Tufts Medical Center, Boston, Massachusetts, USA
[33.] Department of Plastic and Aesthetic Surgery, ATOS Klinik, Munchen, Germany
[34.] Emeritus, Department of Ophthalmology at the University Hospital of the Saarland**,** Homburg, Germany
[35.] Department of Medical Genetics, University Hospital of Antwerp, Antwerp, Belgium
[36.] Department of Otolaryngology and Head and Neck Surgery, Gunma University Graduate School of Medicine, Gunma, Japan
[37.] Center for Chronically Sick Children, Charité University Medicine, Berlin, Germany
[38.] Department of Diagnostic, Interventional and Pediatric Radiology, Inselspital, University Hospital of Bern, Bern, Switzerland.
[39.] Center for Regenerative Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA; Division of Plastic and Reconstructive Surgery, Massachusetts General Hospital, Boston, Massachusetts, USA; Harvard Stem Cell Institute, Cambridge, Massachusetts, USA
[40.] Center for Mendelian Genomics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
[41.] Department of Pathology, University of Iowa Carver College of Medicine, Iowa City, Iowa.
[42.] Department of Human Genetics, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan
[43.] Department of Pediatrics, Cedars Sinai Medical Center, Los Angeles, California, USA
[44.] Medical Genetics, MassGeneral Hospital for Children and Harvard Medical School, Boston, Massachusetts, USA

*****These authors contributed equally to this work.

Correspondence should be addressed to M.E.T. **(**talkowski@chgr.mgh.harvard.edu) or D.R.F. **(**david.fitzpatrick@igmm.ed.ac.uk) or E.E.D. (erica.davis@duke.edu)

Keywords: constraint, methylation**,** pleiotropy, mutation, exome aggregation consortium, hypogonadotropic hypogonadism

ABSTRACT

Arhinia, or absence of the nose, is a rare malformation of unknown etiology that is often accompanied by ocular and reproductive defects. Sequencing of 38 arhinia subjects from 36 independent families revealed that 86% of independent subjects harbor a missense mutation in a constrained ATPase domain of *SMCHD1*. Mutations in *SMCHD1* also cause facioscapulohumeral muscular dystrophy type 2 (FSHD2) via a complex *trans*-acting loss-of-function epigenetic mechanism. Arhinia subjects had comparable DNA hypomethylation patterning to FSHD2 subjects, and CRISPR/Cas9 editing of *smchd1* in zebrafish yielded arhinia-relevant phenotypes. Mutations in *SMCHD1* thus contribute to remarkably distinct phenotypic spectra from craniofacial and reproductive disorders to muscular dystrophy, which we speculate to be consistent with oligogenic mechanisms resulting in pleiotropic outcomes.

Arhinia, or the complete absence of the external nose, is a rare congenital malformation with only 80 patients without holoprosencephaly reported in the past century (see **Supplementary Table 1** for all previous reports**)**. This severe craniofacial dysmorphism can be isolated or accompanied by other craniofacial defects including coloboma, anophthalmia, cataracts, nasolacrimal duct atresia, choanal atresia, and cleft palate (**Fig. 1)**. Seventeen patients with arhinia and ocular defects have been reported with coexistent reproductive failure secondary to hypogonadotropic hypogonadism, a triad called Bosma arhinia microphthalmia syndrome (BAM; OMIM 603457)[1]. In the neonatal period, patients with arhinia are at high risk for respiratory distress, difficulty feeding, and sepsis (as a complication of reconstructive surgery), but those surviving infancy generally demonstrate normal cognitive development with few functional challenges limited to vision loss, the stigma of facial deformities, and osteoporosis and infertility due to hypogonadism. The rarity of these malformations and cross-disciplinary nature of its comorbid conditions have limited systematic efforts to catalog its associated phenotypes, although these comorbidities suggest that genetic factors influencing this condition may have broader developmental implications.

Genetic studies of arhinia have been limited to karyotype analysis, chromosomal microarray, and candidate gene approaches targeting genes related to neural crest cells (NCC) or craniofacial placodal development; to date, no causal locus has been identified. Phenocopies are likewise scarce; homozygous null mutations in *Pax6* arrest nasal placodogenesis in mice[2] and cause rudimentary or malformed noses in humans[3-5]. However, heterozygous and homozygous null mutations in *PAX6* cause aniridia and severe structural brain abnormalities, respectively, that are not observed in individuals with arhinia[3-5]. We formed an international consortium to investigate

the genetic etiology of arhinia and its associated comorbidities and aggregated all available cases across sites. We sequenced 38 individuals with arhinia from 36 independent families as well as 51 family members without ahrinia. Through family-based analyses of *de novo* mutations and genome-wide burden analysis supported by functional studies, we report that rare missense variants in *SMCHD1* represent the predominant single gene contributor to arhinia. Notably, *SMCHD1*, an epigenetic repressor, has also been implicated in a rare, complex oligogenic form of muscular dystrophy (fascioscapulohumeral muscular dystrophy; FSHD2, OMIM 158901). Methylation studies in arhinia patient samples, as well as complementation testing of arhinia variants in a zebrafish model, revealed a common direction of allele effect in both arhinia and FSHD2, a surprising observation considering the striking difference in phenotypes. Given the known oligogenic architecture of FSHD2, these data argue that loss-of-function at the *SMCHD1* locus contributes to the diverse manifestations of arhinia, BAM, and FSHD2, likely through interaction with other genomic loci.

RESULTS

Samples, phenotypes, and epidemiology of arhinia

We established a large international consortium and aggregated all available biospecimens and clinical data to identify the genetic cause of arhinia. This cohort encompassed 24% of all 80 previously reported subjects and an additional 19 new subjects (**Supplementary Table 1**), facilitating a relatively comprehensive picture of the phenotypic spectrum of arhinia (**Supplementary Table 2**). All subjects had complete arhinia, almost universally accompanied by abnormalities of the surrounding craniofacial structures, including high-arched or cleft palate, absent paranasal sinuses, hypoplastic maxilla, nasolacrimal duct stenosis or atresia, and choanal

atresia (**Fig. 1**), and 44% of subjects also had dysmorphic pinnae or low-set ears. Ocular phenotypes included anophthalmia or microphthalmia (69%), uveal coloboma (76%), and cataract (47%), while at least six subjects had normal eye anatomy and vision. Among the 28 subjects in whom the reproductive axis could be assessed (19 male; 9 female), all demonstrated reproductive failure due to hypogonadotropic hypogonadism (HH), and all seven subjects with available brain MRI data presented with absent olfactory structures on imaging; both presentations are hallmark clinical signs of gonadotropin releasing hormone (GnRH) deficiency and anosmia (Kallmann syndrome; OMIM 308700). Twenty-four of these 28 individuals also had ocular defects, indicating that 86% of arhinia subjects that could be assessed met diagnostic criteria for BAM.

Sequencing and gene discovery

To investigate the contribution of rare coding variants to arhinia, we used whole-exome sequencing (WES) in an initial cohort of 22 probands: 9 subjects with available DNA from families of varying pedigree configurations and 13 subjects with no familial samples available (see **Supplementary Fig. 1**). Concurrent with WES, whole-genome sequencing (WGS) was performed in four members of a previously reported multiplex family that included a proband affected with arhinia, an affected sister, an unaffected brother and father, a mother with anosmia and subtle nasal and dental anomalies, a maternal half-aunt with arhinia, and a maternal grandmother who also had mild nasal and dental anomalies (**Supplementary Fig. 1**; family O)[6,7]. These analyses identified rare missense variants in *SMCDH1* in 81.8% of independent probands (**Table 1**), none of which were present in the Exome Aggregation Consortium (ExAC) database of 60,706 healthy individuals with WES[8,9]. Among the eight WES samples with

familial information and an observed *SMCHD1* mutation, we confirmed 3 to have arisen *de novo* and additional 4 samples where the mutation was not observed in available family samples (parent or siblings), while in one subject (family T), the variant was inherited from a father with no craniofacial abnormalities but who carried a clinical diagnosis of muscular dystrophy. To formally test whether the observed allelic distribution in arhinia subjects represented a significant accumulation of rare missense variants, we compared the rare mutation burden among 22,445 genes in the arhinia subjects to the WES data from ExAC (minor allele frequency [MAF] <0.1%). Powered by the size of our aggregate cohort, we found that *SMCHD1* was the only gene that achieved genome-wide significance for a rare mutation burden ($p = 2.9x10^{-17}$, **Fig. 2**).

All of the variants identified in this initial cohort were localized to six of the 48 exons that comprise *SMCHD1* (exons 3, 8, 9, 10, 12, 13; Ensemble transcript ENST00000320876). Based on this narrow distribution, we performed targeted sequencing of these exons in an additional 13 subjects and discovered rare *SMCHD1* missense mutations in 9 of these probands. WES on the *SMCHD1* negative samples identified additional rare missense mutations in three of the four probands remaining, all localized to exons adjacent to the initial six screened in the targeted assays (exons 5,6,11). In these collective analyses (WES, WGS, targeted sequencing), 86.1% (31/36) of independent arhinia probands had a rare missense variant in *SMCHD1* (**Table 1**), and all sporadic subjects with complete trios harbored a *de novo* variant (n = 10). In an additional four multigenerational, multiplex families (O, T, AB, AH) harbored rare missense alleles in *SMCHD1* that segregated with variable phenotypes such as anosmia, asymmetric nares, abnormal dentition, nasal hypoplasia, hypogonadism, and muscular dystrophy, suggesting

incomplete penetrance, variable expressivity, and possible pleiotropy associated with alterations of *SMCHD1* (**Supplementary Fig. 1a)**.

SMCHD1 is among the most highly constrained genes in the genome, suggesting strong intolerance to loss-of-function variation (evolutionary constraint pLI = 1.00)[10], with an estimated combined prevalence of 1 in 10,000 heterozygous null individuals in ExAC. However, the gene does not show particularly strong intolerance to missense variation (81% of expected missense variants observed; p = 0.016). The localization of all arhinia-specific *SMCHD1* mutations to exons 3-13 led us to probe further the distribution of rare missense variants observed in ExAC across this gene. Analyses of regional constraint among the individual exons and critical domains in the protein revealed strong evidence of constraint against missense variation in the 5' region of the gene, including exons 1-19 encompassing an ATPase domain (61% of expected missense variants observed; $\chi^2 = 32.40$; $p = 1.26 \times 10^{-8}$), whereas there was no evidence of in the region including exons 20-48, encompassing an SMC-hinge domain (95% of expected missense variants observed; $\chi^2 = 0.86$; p = 0.36; **Fig. 3**). This observation of strong regional constraint is consistent with the increased burden of rare *SMCHD1* alleles in arhinia subjects, and suggests that these alleles may impede protein function. *In silico* prediction of protein pathogenicity from the Combined Annotation Dependent Depletion (CADD) database revealed that the 19 arhinia-specific *SMCHD1* variants were more deleterious than all rare, nonsynonymous variants in ExAC (MAF < 0.01%, ExAC n = 378, $p = 8.27 \times 10^{-5}$; Supp Figure 2). Importantly, there are 20 rare missense variants in ExAC between exons 3-13 with CADD scores exceeding the median arhinia score (16.91), further supporting our speculation that deleterious *SMCHD1* variant are not fully penetrant, and such variants alone may not be sufficient to manifest arhinia.

Mutational overlap between arhinia and a rare form of muscular dystrophy

SMCHD1 encodes a large protein (2007 amino acids) containing a 5' functional GHKL-type ATPase domain[11] and a 3' SMC-hinge domain (for dynamic DNA binding) that serves as an epigenetic regulator of both autosomal and X-linked genes[12-15]. The discovery of an association between this gene and craniofacial development was unexpected since mutations in *SMCHD1* are associated with FSHD2, a rare oligogenic form of muscular dystrophy. In FSHD2, heterozygous loss of *SMCHD1* repressor activity, in combination with a permissive D4Z4 haplotype on chromosome 4 (4q35), allows for the ectopic expression of the DUX4 protein which is cytotoxic to skeletal muscle[16]. The distribution of mutations in FSHD2 span the entire gene and include missense and truncating variants, whereas all variants observed in arhinia subjects were missense variants clustered tightly around the GHKL-type ATPase domain (**Fig. 3**), which is thought to be critical to the controlled release of DNA bound by SMCHD1[17]. However, we were surprised to find several previously reported FSHD2-specific missense mutations localized to exons 3-13, and one of these FSHD2 variants was also detected in the arhinia cohort (G137E in subject #AG1)[18]. At present, neither subject has features of both disorders, indicating that either these phenotypes have arisen by divergent mechanisms or are influenced by additional loci.

Methylation profiling and protein expression in arhinia and FSHD2 subjects with *SMCHD1* variants

Haploinsufficiency and dominant negative loss-of-function models have both been invoked in FSHD2 for *SMCHD1* mutations that disrupt the open reading frame (nonsense, indel, or splice-site) or preserve it (missense), respectively[18]. In both models, loss of SMCHD1 repressive

activity manifests as a decrease in DNA methylation at SMCHD1 binding sites[11,16,19,20]. Clinical testing for FSHD2 relies on methylation profiling of two of these binding sites, the FSHD-associated 4q35 D4Z4 macrosatellite array and the highly homologous 10q26 D4Z4 array[21,22]. To pursue evidence of mechanistic overlap between arhinia and FSHD2, we quantified 4q35 D4Z4 methylation in 23 arhinia subjects (19 with *SMCHD1* rare missense variants) and 22 family members: 4 with *SMCHD1* rare missense variants, including two with anosmia, one with a hypoplastic nose, and one with symptoms of muscular dystrophy, while the remaining 19 family members were *SMCHD1* mutation-negative. Remarkably, 73.6% (14 of 19) of arhinia subjects with an *SMCHD1* variant had D4Z4 hypomethylation characteristic of FSHD2 (**Fig. 4; Supplementary Table 4**), while all 4 arhinia subjects without a rare missense variant in *SMCHD1* had normal methylation patterns. Two of the four family members harboring an *SMCHD1* variant also displayed D4Z4 hypomethylation, while 17 of the 19 family members without a rare *SMCHD1* variant had normal methylation patterns. These data confirmed that arhinia-specific mutations in *SMCHD1* were associated with the same methylation patterning at D4Z4 as seen in FSHD2, illuminating that two completely distinct phenotypes can arise from alterations to the same genetic locus, and indeed the same alleles, proposing that similar loss-of-function genetic mechanisms may result in distinct phenotypes. We thus turned to *in vivo* modeling to probe the functional impact of *SMCHD1* alterations in animal models.

In vivo **modeling studies of** ***SMCHD1*** **alterations**

To test directly the effect of missense alleles in arhinia patients, and to provide biological evidence for their pathogenicity, we evaluated phenotypes relevant to isolated arhinia and BAM in zebrafish (*Danio rerio*) larvae. While there is no zebrafish structure credibly homologous to

the human nose, facial cartilage patterning is one possible proxy for human craniofacial architecture[23]. Zebrafish eye development is also highly conserved between species, making the zebrafish a robust model for the study of microphthalmia gene candidates identified in human studies[24-26]. *D. rerio* further possesses at least two of the three GnRH paralogs that exist in humans, and the processes by which neurons proliferate, migrate, and maintain the neuroendocrine axis are thought to be largely conserved between humans and teleosts[27-29]. Importantly, the zebrafish genome harbors a single SMCHD1 ortholog (49% identical, 67% similar to human), and the N-terminal encoding the ATPase domain is conserved between the two species (**Supplementary Fig. 2**).

We designed two non-overlapping morpholino (MO) antisense oligonucleotides targeting splice donor sites of two different exons within the *smchd1* genomic region encoding the ATPase domain (e3i3 and e5i5 targeting exons 3 and 5, respectively). The e3i3 or e5i5 *smchd1* MOs (3ng, 6ng, and/or 9 ng/embryo) were injected into embryo batches at the one-to-two cell stage and larvae were evaluated quantitatively for aberrant cartilage patterning, ocular development, and reproductive axis integrity between 1.5 and 3 days post-fertilization (dpf) (**Fig. 5a**). All morphant batches demonstrated a dose-dependent decrease in ethmoid plate width (**Fig. 5a, 5b, Supplementary Fig. 3a**); a dose-dependent increase in ceratohyal arch angle and delayed (or absent) development of ceratobranchial arch pairs (**Fig. 5a, 5c, 5d, Supplementary Fig. 3b-c**) and microphthalmia (tested at a 9 ng dose), all of which demonstrate alterations to phenotypes of relevance to human craniofacial and ear development. Moreover, whole-mount immunostaining of MO-injected embryos with a pan-GnRH antibody followed by ventral imaging revealed a prominent phenotype. The morphant olfactory bulbs and hypothalami were intact; however, the

306 average projection length of the terminal nerve, where GnRH3 neurons reside, was reduced by
307 45% compared with controls ($p < 0.0001$; n = 20 embryos/batch; 2 measurements/embryo,
308 repeated with masked scoring; **Fig. 5g**). The observed cartilage, eye, and GnRH phenotypes
309 were unlikely to be non-specific as each defect was reproduced with both MOs tested, but co-
310 injection of MO and full-length human wild-type (WT) *SMCHD1* mRNA rescued each
311 phenotype significantly (**Fig. 5**). To further confirm these findings and rule out artifacts of MO
312 suppression or toxicity, we targeted the *smchd1* locus using CRISPR/Cas9 genome-editing to
313 generate small insertions and deletions of the coding sequence (**Supplementary Fig. 4**); each of
314 the craniofacial, ocular, and GnRH defects observed in the two morphant models were
315 significantly recapitulated in F0 mutants (Figure 5, **Supplementary Fig. 5**).
316
317 Having established credible quantitative *in vivo* assays, we tested both gain and loss of function
318 paradigms. Injection of either full-length human WT SMCHD1 mRNA or equivalent doses of
319 full-length human mRNA bearing three different recurrent arhinia-associated variants (S135C,
320 L141F, and H348R) into zebrafish embryos independent of MO did not yield appreciable
321 craniofacial phenotypes (**Supplementary Fig. 6**). Moreover, augmented doses of mutant mRNA
322 alone (up to 100 pg) or combinatorial injections of mutant and WT mRNA (100 pg each) was
323 likewise unremarkable arguing that, at least in the context of this assay, a gain-of-function
324 biochemical mechanism is unlikely. Given that suppression of *smchd1* reproduced the three
325 hallmark phenotypes of BAM, we next tested a loss of function paradigm through *in vivo*
326 complementation. Focusing on our most sensitive assay, the quantitatively defined reduced
327 projection of the GnRH-positive terminal nerve, we co-injected either: (1) full-length human WT
328 *SMCHD1* mRNA; (2) human message encoding each of the three variants identified recurrently

in arhinia subjects (S135C, L141F, or H348R); or (3) human message encoding a missense variant (P690S) that causes FSHD2[16] with the e5i5 MO. Full-length human WT *SMCHD1* mRNA, but none of the mutant mRNAs associated with arhinia or FSHD2, rescued the GnRH phenotype (**Fig. 5g**). Complementation of message with a common, and presumably benign, variant from ExAC (V708I; rs2270692) also rescued the phenotype, supporting assay specificity. The likely mode of action of the discovered arhinia alleles in this assay is thus loss-of-function, and we find no foundational differences between the arhinia-specific alleles and the alleles discovered in FSHD2 patients with respect to direction of effect.

Next, we performed CRISPR/Cas9 genome editing on mouse embryos using a guide RNA spanning the boundary between exon 3 and intron 3 of mouse *Smchd1*. The 63 embryos recovered after two zygotic injection sessions displayed a range of variants including WT, homozygous knock-ins of L141F (KI), homozygous knock-outs (KO), compound heterozygotes (L141F/null), and complex compound heterozygous deletions (**Supplementary Table 3**). Unfortunately, multiple attempts to generate non-mosaic L141F heterozygous embryos (WT/KI), akin to what is observed in arhinia subjects, using WT repair templates at equimolar ratio to the mutant repair templates were unsuccessful. Non-mosaic homozygous KO, homozygous KI, and compound heterozygous KO/KI embryos were examined using optical projection tomography[30] at 13.5 days post-conception (dpc) and we observed no morphological or growth anomalies (**Supplementary Fig. 5)**. These results in mouse embryos do not support a simple haploinsufficiency, or indeed null, mechanism causing arhinia but are consistent with previous studies in mice in which *Smchd1* knockdown did not cause craniofacial defects[12,31], though complete knockout of the gene is not viable, and support the notion from the human genetic data

that alteration to a single copy of *SMCHD1* alone may not be sufficient to induce pathology in mammals.

Protein modeling and human expression studies

We investigated the potential impact of these arhinia-specific mutations on SMCHD1 protein structure. The protein structure of the N-terminal region of SMCHD1, where the constrained GHKL-type ATPase domain resides, is unknown. However, the crystal structure of heat shock protein 90 (Hsp90), a member of the GHKL-ATPase protein family found in yeast, is known (PDB: 2CG9), and a recent small-angle X-ray scattering study demonstrated that the ATPase domains of these two proteins are similar in structure[11]. We generated a structural model of the N-terminal region of SMCHD1 with Phyre2[32] (**Fig. 6A)**, with residues mutated in arhinia and FSHD2 highlighted. The top ranking templates identified were Hsp90 structures, covering residues 115-573, although the strongest homology is from approximately residues 120-260. The structural model indicates that the arhinia-specific mutations tend to cluster on the protein surface, suggesting that residues mutated in arhinia may be part of an interaction surface. This hypothesis is independently supported by sequence-based predictors of solvent accessibility (**Fig. 6B**), which reveal a significant tendency for arhinia mutations to be exposed on the protein surface.

Finally, as an initial step towards understanding the pathogenic mechanism of arhinia, we measured SMCHD1 protein levels and performed RNAseq on lymphoblastoid cell lines (LCLs) from 23 total subjects: 10 subjects with arhinia harboring presumably pathogenic *SMCHD1* variants, 11 unaffected family members without *SMCHD1* mutations, and two family members

with a mutation in *SMCHD1* and anosmia or a hypoplastic nose (AH3 and AH5, respectively). In the arhinia subjects, SMCHD1 protein levels appeared to be preserved (on average) in LCLs from subjects with an *SMCHD1* variant compared to controls using two different anti-Smchd1 antibodies (Bethyl A302-872A-M and Abcam ab122555; **Supplementary Fig. 9**). RNAseq analyses incorporated affection status and familial relationships, and differential expression analyses following permutation testing revealed a relatively uniform distribution of p-values compared to expectations. We first compared overall expression changes in SMCHD1 and allele-specific expression differences of SMCHD1 transcripts in arhinia probands. After confirming all mutations in the expressed transcripts that were observed in the DNA analyses, we found that arhinia subjects demonstrated a slight, non-significant decrease in *SMCHD1* mRNA expression compared to controls (fold-change = 0.94, $p = 0.49$), with no average difference in allelic expression of the missense variant compared to the reference allele ($p = 0.70$), indicating no change in message stability in arhinia subjects, at least in the available biomaterials (LCLs).

We next looked for pathways and networks that may be associated with manifestation of arhinia. When considering all differentially expressed genes following permutation testing (unadjusted $p < 0.05$; **Supplementary Table 5**), we discovered alterations to multiple pathways and human phenotypes associated with craniofacial development and epigenetic modification. Remarkably, the strongest human phenotype associated with these genes from ToppGene pathway analysis[33] was "depressed nasal tip" ($p = 5.1 \times 10^{-5}$), as well as related phenotypes such as absent nasal septal cartilage, semilobar holoprosencephaly, small placenta, and median cleft lip and palate ($p < 0.005$). Encouraged by these initial network results across all genes, we next sought greater specificity of these networks by integrating orthogonal chip and RNAseq data generated from

SMCHD1-null mouse neural stem cells (NSCs)[13]. From these analyses, we observed a significant enrichment of down-regulated genes between the human and mouse datasets ($p = 0.029$), but not up-regulated genes ($p = 0.40$), and we identified a high-confidence set of nine overlapping genes that were down-regulated in both datasets; the same phenotype, "depressed nasal tip", was more significant than in the human data alone ($p = 2.1 \times 10^{-5}$). We found these results, and multiple related human phenotype associations, to be primarily driven by two genes (*TGIF1*,*DOK7)*, both of which have already been demonstrated to play a role in craniofacial morphogenesis. *DOK7* haploinsufficiency causes fetal akinesia deformation sequence (FADS; OMIM 208150). The phenotypic spectrum of this disorder includes many features of BAM (depressed nasal bridge, cleft palate, choanal atresia, microphthalmia, cataract, coloboma, cryptorchidism, and absent olfactory structures)[34], and heterozygous loss-of-function mutations in *TGIF1* cause holoprosencephaly-4 (OMIM 142946), which may include arhinia, microphthalmia, and cleft palate[35]. Notably, two additional genes associated with the "depressed nasal tip" phenotype, *ICK* and *KDM6A* (an X-linked gene previously associated with Kabuki syndrome, including multiple craniofacial anomalies and cleft palate)[36,37] were also differentially expressed in the human dataset. These four genes are therefore rational mechanistic candidates for modifiers of the arhinia phenotype in the presence of *SMCHD1* mutations.

Prediction of phenotypic outcomes

Among the most striking findings in this study was the demonstration that variants in the same 5' constrained region of SMCHD1 are associated with both FSHD2 and arhinia, and the discovery of an identical amino acid substitution (G137E) within this region that was associated with both phenotypes[18]. Methylation assays suggested indistinguishable hypomethylation signatures

between arhinia and FSHD2 probands. To our knowledge, the comorbid presentation of arhinia and FSHD2 has never been reported. The lack of previously reported FSHD2 symptoms in arhinia subjects may be a consequence of the oligogenic architecture of FSHD2, which would suggest that only a small subset of subjects with arhinia, an already rare condition, would harbor the requisite genetic architecture at D4Z4 and thus be both at risk and past the average age at onset for FSHD2. In addition, features such as facial weakness, which is often one of the first clinical signs of FSHD2, could easily be overlooked or dismissed in a patient with craniofacial anomalies who has undergone multiple corrective surgeries. Nonetheless, we performed analyses comparable to clinical diagnostic testing and found 2 arhinia probands (A1 and E1) with *SMCHD1* variants who met the four critical clinical criteria for susceptibility to FSHD2: 1) an *SMCHD1* pathogenic variant, 2) D4Z4 hypomethylation (bisulfite sequencing [BSS] <25%), 3) a permissive chromosome 4q haplotype, and 4) an 11-28 D4Z4 repeat unit at the 4q array (**Supplementary Table 4**)[16,21,38,39]. Five other subjects may be at risk for FSHD2 but will require additional confirmatory clinical testing. We had consent to re-contact both arhinia subjects meeting FSHD2 clinical criteria, and phenotypic evaluation suggested that at least one subject had symptoms of FSHD2, proposing yet another mutation site (N139H) common to these two disorders. Overall, these results suggest that at least two mutations (G137E and N139H), in the presence of a specific genetic background, can manifest as two divergent clinical phenotypes.

DISCUSSION

We describe genetic, genomic, and functional evidence that implicate *SMCHD1* as the predominant driver of arhinia in humans. These analyses represent the first evidence of a genetic cause for this rare craniofacial malformation. Through a large collaborative effort, we were able

to combine data from a sizeable fraction of subjects reported in the literature (24%) and 19 new subjects, which facilitated the uniform evaluation of the clinical phenotype associated with this condition. We find that 86% of subjects with arhinia who could be assessed present with the BAM triad, and that 88% of subjects with BAM harbored *SMCHD1* variants. In addition, the three BAM subjects without an *SMCHD1* variant were either part of a consanguineous family, or exhibited unique phenotypic features (e.g., tracheoesophageal fistula) suggesting alternative genetic causes in these individuals. Our findings thus suggest a novel role for *SMCHD1* in cranial NCC migration and/or craniofacial placodal development.

Our genetic observations raise questions concerning potential molecular mechanisms that lead mutations in the same gene to produce the distinct phenotypes. The fact that all of the arhinia-associated mutations are missense changes rather than truncating mutations, as often seen in FSHD2, suggests that the arhinia mechanism in humans requires production of a mutant protein rather than simple loss of function of one allele seen in FSHD2. However, the overlap of some arhinia mutations with missense alterations observed in FSHD2 suggests that the mutant protein that is produced is indeed deficient in some critical function. The FSHD2 hypomethylation signatures associated with *SMCHD1* mutations are demonstrably loss-of-function and consistent in most circumstances with a haploinsufficiency model, although dominant-negative activity of the mutant protein has been suggested as the cause of a more severe phenotype in some cases[16,18,20]. We find largely identical methylation patterns at the D4Z4 repeat region on chromosome 4 in arhinia probands and FSHD2 patients, supporting the view that loss of this function of *SMCHD1* also occurs in arhinia, and so does not in itself explain the difference in phenotypic outcome. Thus, additional factors must be involved in producing this distinction,

such as interaction at the genetic level with variants at other loci or a function-altering interaction at the protein level of the mutant SMCHD1 protein. Indeed, we found that the arhinia-specific variants tend to cluster on the surface of the protein, potentially facilitating disruption of interactions with protein partners, either wild-type SMCHD1 or other members of its complexes (or both). Correspondingly, we found no significant difference in average protein expression between arhinia probands and unaffected individuals, suggesting that the bioactivity of the protein is the critical factor in humans rather than the total amount of protein.

The distinct findings in two model systems reinforce the complexity suggested in humans. The zebrafish model supports the involvement of loss of function as both MO suppression and mosaic ablation of *smchd1* result in BAM-related phenotypes, the most dramatic of which is the GnRH terminal nerve projection defect. These results are specific, as these phenotypes are rescued with full-length human *SMCHD1* WT mRNA, but not mRNA containing recurrent arhinia mutations, and substantial overexpression confers no discernible phenotype. In the mouse, complete loss of function has been achieved as homozygosity for an exon 23 *Smchd1* nonsense mutation which produces hypomethylation and mid-gestational lethality in females[12], although males are viable. Heterozygosity for this mutation, like induction of either deletions or arhinia-relevant point mutations by CRISPR/Cas9, produced no phenotypes in mouse. Unfortunately, we were unable to replicate in the mouse the heterozygous missense genotype characteristic of human arhinia (**Supplementary Fig. 5**). The fact that loss of smchd1 is sufficient produce BAM-relevant phenotypes in the zebrafish, but loss of Smchd1 in the mouse does not, reinforces the need to consider genetic and functional interactions of the mutant protein in causing the human arhinia phenotype. Notably, both the zebrafish and mouse genomes lack

490 recognizable orthologs of *DUX4*, the genetic interactor necessary for the development of
491 FSHD2[40,41].
492
493 The complex oligogenic architecture of FSHD2 suggests that only a small fraction of individuals
494 with arhinia, which is exceedingly rare on its own, will have an *SMCHD1* mutation and also
495 carry a permissive 4q35 haplotype, placing them at risk for FSHD2. Our analyses identified
496 seven subjects that are potentially at risk for FSHD2, and at least one appears to display
497 symptoms of the disorder. Nonetheless, one-quarter of individuals who meet genetic diagnostic
498 criteria for FSHD2 are clinically asymptomatic, indicating that the full complement of genetic
499 requirements for developing clinical FSHD2 is not yet known[42]. Like our data, the absence of
500 arhinia in patients with FSHD2 with *SMCHD1* mutations within the constrained ATPase domain
501 argues that loss of SMCHD1 activity alone is not sufficient to produce a craniofacial phenotype.
502 The same is true for patients with FSHD2 with mutations in DNA methyltransferase type 3B
503 (*DNMT3B*) [43] who have no clinical signs of immunodeficiency, centromeric instability, and
504 facial anomalies syndrome type 1 (ICF1 [OMIM: 242860])[44], the autosomal recessive disorder
505 associated with mutations at this locus. Within our cohort, we observed multiple family members
506 harboring *SMCHD1* mutations with only mild dysmorphic features or anosmia, and at least one
507 individual without any dysmorphic features. Given the epigenetic function of *SMCHD1*, it is
508 plausible that one or more genetically-interacting loci influence susceptibility to arhinia with the
509 proximal interactors of *SMCHD1* such as *TGIF1* and *DOK7* representing prime candidates.
510 Disentangling these genetic mechanisms in conjunction with the biochemical consequences of
511 *SMCHD1* missense mutations in humans and model organisms will be a critical area of further

study, ideally in human tissue of relevance to arhinia and FSHD2 rather than the LCLs currently available.

In conclusion, we discovered that rare variants localized to an evolutionarily constrained region of *SMCHD1* are associated with BAM and isolated arhinia. Importantly, during the course of this study we learned of an independent effort by Gordon and colleagues, who also identified *SMCHD1* missense mutations in arhinia subjects. In correspondence we have compared our subjects and determined that their study provided an additional 7 subjects that were independent of our analyses (six overlapped), bringing the total to 45 arhinia subjects, (87%) of whom harbored a rare missense mutation in *SMCHD1*. Their analyses also confirmed the *SMCHD1* mutation to have occurred *de novo* in two of the overlapping subjects for which we did not have parental samples (M1 and AJ1). The molecular mechanism by which such mutations contribute to arhinia, and what differentiates FSHD2 and arhinia patients, remains unclear, though our functional modeling suggests that a simple, single locus mechanism is unlikely. Our analyses thus emphasize yet another example in a growing list of genes in which mutations can give rise to pleiotropic phenotypes across the spectrum of human anomalies. For *SMCHD1*, these phenotypes – a rare muscle disease and now, a severe craniofacial and reproductive disorder - are perplexingly diverse. Dissecting the genetic and epigenetic factors that determine phenotypic manifestations will inform both our understanding of the pathogenesis of the arhinia-BAM-FSHD spectra and, more broadly, the genetic and epigenetic architecture of oligogenic disorders.

ACKNOWLEDGMENTS

We thank all participants, family members, and clinical staff for their generous contributions of time and materials to this research. We thank Tammy Gillis, Jayla Ruliera, and Mary Anne Anderson for technical assistance. This project was funded by grants from the National Institutes of Health (R00MH095867 to M.E.T., P01GM061354 to M.E.T., J.F.G., C.C.M., and E.C.L., T32HD007396 to H.B., R01HD081256 to M.E.T., P50HD028138 to W.F.C, S.B.S., M.E.T., N.K., and E.E.D., K23HD073304-02 and 1SI2ES025429-01 to N.D.S., R01AR062587 to P.L.J.), the March of Dimes (FY15-255 to M.E.T.), the Medical Research Council (MR/M02122X/1 to J.A.M.), the German Research Foundation (SFB665 to A.M.K), the Berlin Institute of Health (BIH-CRG1 to A.M.K.), D.R.F, He.B, K.A.W, J.R, J.K.R and J.A.M are funded by program grants from the Medical Research Council (MRC) Human Genetics Unit award to the University of Edinburgh. MA is funded by the University of Edinburgh Institute of Genomics and Molecular Medicine Translational Intitiative Fund. N.K. is a distinguished Jean and George Brumley Professor, and M.E.T. is supported as the Desmond and Ann Heathwood MGH Research Scholar.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

AUTHOR CONTRIBUTIONS

M.E.T., D.R.F., E.E.D., N.K., P.J., N.D.S., and H.B. designed the study. N.D.S., L.P., K.A.W., M.N., S.P., T.K., D.L., A.S., S.J., J.C.S., M.F.L., S.S., N.P., J.L., N.F., A.V., A.R., K.S., I.S., D.S., N.O., C.J., J.T., S.C., L.A.S., B.B., C.C., J.E.G., T.P.B., O.P.S., J.D.H., W.M., K.W.R.,

B.L., M.S., A.K., C.H.C., C.C.M., V.V.H., R.B., J.E.H., S.B.S, J.M.G., A.E.L., W.F.C., D.R.F., recruited patients and collected clinical information and samples. Z.A.K, He.B., L.P., S.E, T.I.J., J.R.W, C.G, A.S., C.M.S, Y.A, B.B.C., M.A., J.K.R, M.Z., J.W.J., E.L., S.A.M., N.K, P.L.J, E.E.D, D.R.F performed molecular genetic and animal modeling studies, H.B., K.S., R.L.C, A.L, M.L, J.F.G, D.G.M, M.E.T performed genomic analyses, J.M. performed protein modeling. N.D.S., H.B., N.K., J.F.G., P.L.J., E.E.D, D.R.F., and M.E.T. wrote the manuscript, which was revised and approved by all co-authors.

SUPPLEMENTARY INFORMATION

Supplementary material is available online and contains Supplementary Figs. 1-8 and Supplementary Tables 1-4.

Table 1. *SMCHD1* mutations observed in arhinia cohort

Chr	Nucleotide change	Exon	Inheritance (Sample ID)	# Subjects	Sample IDs	AA Change	Gender (Sample ID)
18	c.2666926T>C	3	N/A	1	K1	p.L107P	F
18	c.2666992T>A	3	N/A	1	D1	p.M129K	M
18	c.2667009A>T	3	De Novo (AF1) N/A (M1)	2	M1,AF1	p.S135C	F
18	c.2667010G>A	3	De Novo (I1) N/A *(R1)	2	I1, R1	p.S135N	F(R1), M(I1)
18	c.2667014A>C	3	Father*	1	T1	p.E136D	M
18	c.2667016G>A	3	N/A	1	AG1	p.G137E	F
18	c.2667021A>C	3	De Novo (A1) N/A (Y1)	2	A1,Y1	p.N139H	F(A1,Y1)
18	c.2667029G>C	3	N/A	3	C1,E1,S1	p.L141F	F(S1), M(C1,E1)
18	c.2667029G>T	3	De Novo	1	V1	p.L141F	M
18	c.2674017 T>G	5	N/A*	1	AB1	p.F171V	M
18	c.2688478C>G	6	De Novo	1	AA1	p. A242G	M
18	c.2694685A>G	8	Mother*	2	O1, O4**	p.Q345R	F
18	c.2697032A>G	9	De Novo (X1,AC1,AE1) N/A (F1,L1,N1,Z1)	7	F1,L1,N1,Z1 X1,AC1,AE1	p.H348R	F (L1,X1), M(F1,N1,Z1, AC1,AE1)
18	c.2697896A>T	10	Father*	1	AH1	p.Q400L	F
18	c.2697956A>T	10	De Novo	1	P1	p.D420V	M
18	c.2700611G>C	11	N/A	1	W1	p. E473Q	M
18	c.2700837C>A	12	N/A	2	J1,U1	p.T523K	F(U1), M(J1)
18	c.2700840A>G	12	N/A	1	B1	p.N524S	M
18	c. 2703697G>A	13	N/A	1	AJ1	p.R552Q	M

*Multiplex family

**Siblings

A rare missense mutation was not identified in *SMCHD1* in subjects G1, H1, H2, Q1, AD1, or AI.

N/A = parental samples not available; AA = amino acid; M = male; F = female.

Amino acid codes: A=Ala, R=Arg, N=Asn, D=Asp, C=Cys, Q=Gln, E=Glu, G=Gly, H=His, L=Leu, M=Met, F=Phe, P=Pro, S=Ser, T=Thr, V=Val.

FIGURE LEGENDS

Figure 1. Phenotypic spectra associated with arhinia

Five representative subjects (a-e) demonstrating complete congenital arhinia and variable ocular phenotypes: a) Subject V1 (age 2) with left-sided iris coloboma b) Subject AC1 (age 10) has left-sided microphthalmia and bilateral nasolacrimal duct stenosis c) Subject U1 (as a newborn) has normal eye anatomy and vision, d) Subject O4 (age 16) has right-sided microphthalmia e) Subject A1 (young child, age unknown) has bilateral colobomatous microphthalmia, cataracts, and nasolacrimal duct atresia. f-j) All craniofacial radiographic images are from subject V1: f) Surface rendering reconstruction from a MRI 3D T1 weighted sequence showing complete absence of the nose (arrow1) g) 3D volume rendering technique (VRT) reconstruction from spiral CT showing complete absence of nasal bones (arrow) h) Coronal reconstruction from CT showing absence of nasal septal structures. The maxilla articulates with the nasal process of the frontal bone (arrow) i) Coronal MRI T2 weighted sequence showing absence of the olfactory bulb and olfactory sulcus (arrow) j) Midline MRI sagittal T1 weighted sequence. There is a high-arched palate (cleft not visible on this image) and decreased distance between the oral cavity and the anterior cranial fossa (black arrow). The rudimentary nasopharynx (*) is blind and air-filled. The pituitary gland (white arrow) appears normal.

Figure 2. Association analyses for rare mutation burden in arhinia

Manhattan plot and quantile-quantile (q-q) plot demonstrating the significant accumulation of rare *SMCHD1* mutations in subjects with arhinia compared to the ExAC cohort ($p = 2.9e-17$). Analyses involved a variant count at each gene for arhinia subjects compared to ExAC controls ($n = 60,706$) who presumably do not have arhinia after filtering for allele frequency (MAF <

606 0.1%), quality (mean depth ≥10; mapping quality ≥10) and predicted function (nonsynonymous,
607 splice site, and frameshift mutations). Any gene with at least one mutation passing these criteria
608 was included in the analysis (n = 22,445 genes). Genome-wide significance threshold was $p < 2.2x10^{-6}$
609 following Bonferroni correction (red line) and only *SMCHD1* achieved this threshold.
610
611
612 **Figure 3. Arhinia mutations occur near the 5' GHKL-type ATPase domain**
613 a) The distribution of arhinia mutations across *SMCHD1* is tightly clustered between exons 3-12
614 of the gene compared to b) the distribution of variants observed in *FSHD2* subjects and c) ExAC
615 controls. *FSHD2* subjects were taken from the Leiden Open Variation Database (LOVD 3.0)[45].
616 Constraint analysis as described by Daly and colleagues[10] revealed that while the gene displays
617 significant overall intolerance to deleterious mutations ($p = 0.016$), this significance is almost
618 entirely driven by constraint across the first 19 exons of *SMCHD1* ($X^2 = 37.73$; $p = 8.12x10^{-10}$),
619 which includes the GHKL-type ATPase domain, whereas the region from exons 20-48 are not
620 constrained ($X^2 = 0.87$; $p = 0.35$). Figures were modified from the cBioPortal Mutation Mapper
621 software v1.0.1 (http://www.cbioportal.org/mutation_mapper.jsp)[46,47]
622
623 **Figure 4. DNA methylation analysis of D4Z4 repeats**
624 a) Bisulfite sequencing (BSS) of the chromosome 4q and 10q D4Z4 repeats identifies DNA
625 hypomethylation consistent with dominant *SMCHD1* hypomorphic mutations found in FSHD2
626 patients. A total of 52 CpGs were analyzed, arranged linearly from left to right, for 12 clones
627 arranged top to bottom, each representing an independent chromosome analyzed. Each predicted
628 CpG is represented by a box, with red boxes indicating methylated CpGs and blue boxes
629 indicating unmethylated CpGs. b) Cartoon of the chromosome 4q and 10q D4Z4 macrosatellites

that vary in repeat units (RU) from 1 to ~120 RUs. The region analyzed by BSS in each RU is indicated by a green bar. *FSHD2 requires a mutation in *SMCHD1* combined with at least 1 chromosome 4q D4Z4 array ranging in size between 11-28 RUs and a permissive A-type 4q subtelomere. c) BSS observed 75% of arhinia probands with *SMCHD1* mutations that could be tested due to available material had D4Z4 hypomethylation characteristic of FSHD2, while the single proband tested without a *SMCHD1* mutation showed a normal methylation pattern. BSS was measured from the lowest quartile as previously described[21] and a methylation rate of <25% was considered consistent with hypomethylation observed in FSHD2. See **Supplementary Table 4** for further details on individual methylation status.

Figure 5. *In vivo* modeling of *smchd1* in zebrafish demonstrates craniofacial and GnRH phenotypes relevant to congenital arhinia

a.) Suppression of *smchd1* results in altered cartilage structures in 3 day post-fertilization (dpf) -*1.4col1a1:egfp* larvae. Representative ventral images; *smchd1* morphants and F0 mutant larvae display a reduction in the size of the ethmoid plate (ep, as measured with solid white arrows); and abnormal jaw structures including a broadened ceratohyal angle (ch, dashed white line), and reduction in the number of ceratobranchial arches (cb, asterisks). Scale bar, 200μm b) Quantification of ethmoid plate width measured on ventral images. The furthest distal width (a, left panel a) was normalized to the width at the ethmoid plate-trabecula junction (b, left panel a). c) Loss of *smchd1* results in a decreased eye size; lateral bright-field images of representative 3dpf control, morphant, and CRISPR/Cas9 larvae are shown. Scale bar = 300μm. d) Quantification of eye size area in larval batches (indicated with dashed white circle in panel c). e) Immunostaining of gonadotropin releasing hormone (GnRH) neurons in 1.5 dpf embryos with

a pan-GnRH antibody shows shorter terminal nerve (tn) projections from the olfactory bulb (ob) in *smchd1* models. Representative ventral views are shown; h, hypothalamus; scale bar, 100μm; Dashed white boxes are zoomed to show tn projections in the insets; dashed white lines (insets) indicate tn length measurement starting proximal to the ob, and extending to the tip of the tn. f) *In vivo* complementation assay of missense *SMCHD1* variants using GnRH tn length as a phenotypic readout. S135C, L141F, and H348R are recurrent mutations in arhinia cases; P690S is associated with FSHD2[18] V708I (rs2276092) is a common variant in ExAC and is a negative control for the assay. Orientation indicated (panels a and e) with arrows pointing to anterior (A), posterior (P), left (L) and right (R). Statistical significance is indicated with *** ($p<0.0001$), ** ($p<0.01$), or *($p<0.05$); g, guide RNA; NS, not significant. n=19-50 embryos/injection (panel b); n=28-59 embryos/injection (panel d); n=18-20 embryos/injection (panel f) with masked scoring; all experiments were repeated. Error bars indicate standard error of the mean.

Figure 6: SMCHD1 protein modeling.

Protein modeling predicts that arhinia mutations were more likely to occur on the surface of Smchd1 and disrupt a binding surface compared to the distribution of FSHD2 mutations. A) Homology model of the N-terminal region of SMCHD1 generated with Phyre2[32] with residues mutated in arhinia (red) and FSHD2 (blue). All of the top 20 structural templates had GHKL domains: 16 were Hsp90 structures, two were mismatch repair proteins (MutL/Mlh1) and two were type II topoisomerases. Only those residues modeled with high confidence are shown (115-295; 314-439; 458-491; 504-535; 552-573). B) Comparison of predicted relative solvent accessibility values for residues in the N-terminal region of SMCHD1 mutated in arhinia and FSHD2. Three different predictive methods were used: NetsurfP[48], I-TASSER[49] and SPIDER[50].

Residues mutated in both disorders (136-137) are excluded in this analysis. P-values are calculated with the Wilcoxon rank-sum test. Boxes represent quartile distributions.

ONLINE METHODS

Research Subject Enrollment. We collected existing DNA or blood samples from 38 subjects with arhinia (22 male, 16 female). Whenever possible, DNA was also collected from family members. Phenotypic information was obtained via questionnaires completed by patients, parents, or referring physicians and confirmed by review of official medical records and consultation with the referring physician. Note that reproductive axis dysfunction could not be determined in pre-pubertal girls or in pre-pubertal boys without congenital microphallus or cryptorchidism. All research was approved by the Institutional Review Board of Partners Healthcare and a subset of families consented to publication of photographs (**Figure 1**).

Whole-Exome Sequencing (WES). We performed WES on 26 total probands with arhinia (22 in initial round and 4 that failed targeted sequencing) and 12 family members. The majority of subjects (n = 29) were sequenced at the Broad Institute (Cambridge, MA, USA), including 21 independent subjects and 1 set of affected siblings from a consanguineous family. We also sequenced 6 unaffected available family members from these subjects at the Broad Institute (families A, D, E; see **Supplementary Fig. 1**). We collected another two sporadic subjects, one trio (family V) and a mother-proband pair (family U), that had previous WES sequencing from the University of Zurich (Zurich, Zurich, Switzerland). We also collected a trio (family T) that had previously undergone WES by GeneDx (Gaithersburg, MD, USA) and contained an affected proband who also had a deceased great aunt with arhinia and coloboma. We finally received exome results for a subject (AJ1) with arhinia from the Department of Human Genetics at Nagasaki University. All exomes except sample AJ1 were aligned in house with BWA-MEM v.0.7.10 to GRCh37 and underwent joint variant calling by GATK[51] following best practice methods[52,53]. Familial relationships were confirmed by KING v1.4[54] and variants were annotated with Annovar v.2016-02-01[55] against the refseq annotation of the genome (http://www.ncbi.nlm.nih.gov/refseq/).

Whole-Genome Sequencing (WGS). We obtained samples from 4 members of multigenerational family O [6,7] (see Supplementary Fig.1) and performed whole-genome deep WGS to 30X average coverage on the Illumina X Ten platform. Family O had multiple individuals with craniofacial abnormalities beyond the proband's arhinia, including a deceased maternal-half aunt with arhinia, a sister with arhinia, a mother with anosmia and subtle nasal and dental anomalies, and a maternal grandmother with mild nasal and dental anomalies. Note that samples from the affected sister, unaffected brother, and unaffected maternal half-aunt were obtained after WGS had been completed and were therefore screened for the p.Q345R variant by targeted sequencing. Variants were aligned with BWA-MEM v.7.7 to GRCh37 and GATK was used to call single nucleotide variants (SNVs) as described above.

Genetic Association Analyses. We compared the genic burden of rare, nonsynonymous variants detected by WES in independent arhinia subjects from our cohort (n = 29; one affected subject [brother] selected from consanguineous sibship) with WES data from over 60,706 controls in the the Exome Aggregation Consortium[8,9] (ExAC; http://exac.broadinstitute.org/). Analyses were restricted to include variants that passed the following criteria: 1) high quality (GATK Filter=PASS), 2) rare (ExAC minor allele frequency [MAF] < 0.1%), 3) mean depth ≥ 10 reads, 4) a mapping quality ≥ 10, and 5) predicted to be nonsynonymous, to alter splicing, or to cause a frameshift. As there was no gender bias among our arhinia subjects to suggest sex-linkage (42% female), and we could not ascertain gender from the ExAC database, analyses were restricted to autosomes. Counts between ExAC and the arhinia cohort were compared by a Fisher exact test. Results were visualized as a Manhattan and QQ plot created by the R package qqman[56].

Targeted Sequencing. Variants of interest, as determined by our WES and WGS gene association analysis, were subsequently confirmed by Sanger sequencing in all subjects except T1, as DNA was not available (we are getting DNA). Analyses of these subjects demonstrated a significant aggregation of rare mutations in *SMCHD1* restricted to exons 3, 8-10, 12, and 13. We therefore performed targeted sequencing of these exons in all additional subjects (n = 12) using the primers below and subjects that failed this targeted sequencing (n=4) were sent for WES as described above.

Exon	Primer Sequence 5'-3'
Exon 3 fwd	TGCTTACAGGTAGATGATTGGG
Exon 3 rev	GGAATGGGATACGTAATCAGG
Exon 6&7 fwd	TTAACACTGAATACAAGTGCAATG
Exon 6&7 rev	TTCATACTTTCAAGTTAAGTTCTGTCC
Exon 8 fwd	TGTATTGGGCCAGTTTCCTC
Exon 8 rev	CCTGTGCCTCAAATAATGCTC
Exon 9 fwd	AAATGCTTAATAAAGTGCTTGATACC
Exon 9 rev	TTTATTATCCTGAGTCATTTGGAAC
Exon 10 fwd	TGTCCTTCAGCTCTGATTTGC
Exon 10 rev	GAGAAGACAAGGGAACATATAAAGG
Exon 11 fwd	TGTGTTTGTTTCATTATTTCTCACA
Exon 11 rev	GGAGGAGTACACCAGTCAAAGC
Exon 12 fwd	CAGCTAGAGGGAAAAGGCCT
Exon 12 rev	TGTGAACACTTGACTGCTCA
Exon 13 fwd	GGTAATGCATTTGTTTGAAATATCC
Exon 13 rev	CTTCATGAAATGTGAGAATGGG

Inheritance Testing: For samples with a predicted de novo variant without WES we confirmed familial relationships by determining repeat length of 10 STS markers (d15s205, d12s78, d4s402, d13s170, d4s414, d22s283, d13s159, d2s337, d3s1267, d12s86). Inheritance of markers was checked in each proband and proper parental inheritance was confirmed in all cases.

Inheritance for a single proband (P1) was confirmed in a similar manner at the University of Edinburgh with the following nine markers: cfstr1, d7s480, dxs1214, amel, nr2e3_22, d4s2366, i1cahd, d5s629, d5s823.

Transcriptome Sequencing (RNAseq). Total RNA of ~1 million cells was extracted from EBV-transformed lymphoblastoid cell line (LCLs) using TRIzol® (Invitrogen) followed by RNeasy® Mini Kit (Qiagen) column purification. RNAseq libraries were prepared using the Illumina TruSeq kit and manufacturer's instructions, as described[57,58]. Libraries were multiplexed, pooled and sequenced on multiple lanes of an Illumina HiSeq2500, generating an average of 33 million paired-end reads of 76 bp. Quality checking of sequence reads was assessed by fastQC (v. 0.10.1) (**http://www.bioinformatics.babraham.ac.uk/projects/fastqc/**). Next, sequence reads were aligned to human reference genome Ensembl GRCh37 (v. 71) using GSNAP (v. 12-19-2014) at its default parameter setting[59]. Quality checking of alignments was assessed by a custom script utilizing Picard Tools (**http://broadinstitute.github.io/picard/**) , RNASeQC[60], RSeQC[61] and SamTools[62]. Gene level counts were tabulated using BedTools's multibamcov algorithm (v. 2.17.0)[63] on unique alignments for each library at all Ensembl genes (GRCh37 v.71). We found the threshold to detect expressed genes to be at least six uniquely mapped reads by relying on analysis of External RNA Controls Consortium (ERCC) spike-ins as we have previously described[57]. After filtering out short genes (transcript lengths < 250 nt) and rRNA and tRNA genes, only the 15,936 genes that met the detection threshold in all case samples or all control samples were kept for further analysis. To account for the effect of the covariance among family members, a generalized linear-mixed model (GLMM) approach was used. For this task, a mixed model package, lme4 (v. 1.1.10)[64] was employed in R (v. 3.2.2). Specifically, gene-level expression data across samples as raw counts was fitted to a following GLMM based on a Poisson-lognormal approach $condition + (1|familyId) + (1|obsId)$, where *condition* is a fixed factor that describes a binary disease status of an individual, *familyId* is a random factor that accounts for similarity in expression due to shared genetic background and *obsId* is a random factor that accounts for individual-level random effects. This model converged on 15,478 genes. An evolutionary constrained gene list was retrieved from the ExAC database (v. 0.3 release 3-16-2015), where constrained genes were defined to be those with a probability of being intolerant to loss of function mutations ≥ 0.9. A protein-protein interaction network of differentially expressed genes (nominal $p < 0.05$) was constructed based on physical interaction data from the BioGRID database (v 3.4.135)[65]. The resulting network contained 1,069 proteins and 2,593 pair-wise interactions in which a protein had 4.86 connections (degrees) on average. We defined hub proteins to be in the top 5th percentile of degree distribution in this network, which corresponds to 17 connections or more.

Western Blot: Protein was harvested from 1 million LCLs in 23 total subjects: 10 subjects with arhinia harboring presumably pathogenic *SMCHD1* variants, 11 unaffected family members without *SMCHD1* mutations, and two family members with a mutation in *SMCHD1* and anosmia or a hypoplastic nose (AH3 and AH5, respectively; **Supplemental Fig. 9**). Protein extraction was performed with the following procedure: 1) Cells were washed in 1x PBS and lysed in 300 ul ice-cold 1 x RIPA buffer (http://www.bio-world.com/productinfo/4_62_465/7465/RIPA-Buffer-X-pH.html) supplemented with 5 mM PMSF. 2) After 30 min. incubation on ice, cell lysates were cleared by centrifugation (15G, 15 min., 4*C) and soluble proteins concentration was assayed with BCA reagent

(**https://www.thermofisher.com/order/catalog/product/23225#/23225**). Extracted proteins (15-30 ul/sample) were next separated by a 8% sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE;Bio-Rad MiniProtean 3 Cell, 2 hr 15 mA) and transferred onto a polyvinylidene fluoride (PVDF) membrane (Bio-Rad cat#1620174) using liquid transfer system (Bio-Rad Ready Gel Cell) at 4*C, 10V for 16 hrs. Western plotting was performed using two sets of SMCHD1 antibodies: 1) Bethyl Laboratories A302-872A-M (anti-SMCHD1, C-terminus); 2) Abcam ab122555 (anti-SMCHD1, N-terminus). We used two loading control antibodies: 1) Abcam ab6046 (beta-Tubulin load control) 2) Abcam ab8227 (beta-Actin load control). Antibody dilutions were used as recommended by manufacturer. Primary antibodies were diluted in tris-buffered saline and tween 20 (TBST) buffer and 1% BSA, secondary HRP-conjugated antibody (1:20,000 dilution) in TBST without BSA. Membrane was cut alongside 75 kDa marker (BioRad Precision Plus Protein standards cat# 161-0375) and the upper part was used for blotting SMCHD1 (MW=250 kDa), while the lower part for blotting beta-Tubulin (MW=50 kDa) and beta-Actin (MW=42 kDa) controls. Blotting with primary antibody was carried out overnight at 4*C on a rocking platform, followed by three 10 min. washes in TBST at room temperature. Blotting with secondary antibody was carried out at room temperature for 1 hr, followed by three 10 min. washes in TBST. Re-blotting of SMCHD1 with an alternative antibody, the previously used primary antibody was stripped off with mild stripping buffer, as described: http://www.abcam.com/ps/pdf/protocols/stripping%20for%20reprobing.pdf. Western blot were luminesced with ECL reagent (Bio-Rad cat# 170-5060) and developed with the ChemiDoc MP system (http://www.bio-rad.com/en-us/product/chemidoc-imaging-systems/chemidoc-mp-system). Automated protein quantification was done using Image Lab 5.2.1 software (BioRad).

CRISPR/Cas9 Genome Editing in Mouse Embryos. To generate mouse embryos carrying the p.Leu141Phe disease associated missense variant in *Smchd1*, a double stranded DNA oligomer (CCTTTGCGTAAGTAACCTGCTC) that provides a template for the guide RNA sequence was cloned into px461. The full gRNA template sequence is amplified from the resulting px461 clone using universal reverse primer and T7 tagged forward primers. The guide RNA PCR template is used for *in vitro* RNA synthesis using T7 RNA polymerase(Neb), and the RNA template is subsequently purified using RNeasy mini kit (Qiagen) purification columns. Cas9 mRNA was procured from Tebu Bioscience. The wild-type and mutant repair templates (chr17:71,463,705-71,463,818 GRCm38) are synthesized as 114bp ultramers bearing the desired sequence change from IDT. The injection mix contains Cas9 mRNA (50ng/ul), guide RNA (25ng/ul) and repair template DNA (150ng/ul). Injections are performed in mouse zygotes and the embryos are later harvested for analysis at 11.5 and 13.5 dpc stage of embryonic development.

Optical Projection Tomography. Whole mouse embryos were mounted in 1% agarose, dehydrated in methanol and then cleared overnight in BABB (1 part Benzyl Alcohol: 2 parts Benzyl Benzoate). The sample was then imaged using a Bioptonics OPT Scanner 3001 (Bioptonics, UK) using tissue autofluorescence (excitation 425nm/emmision 475nm) to capture the anatomy. The resulting images were reconstructed using Bioptonics propriatory software, automatically thresholded and merged to a single 3D image output using Bioptonics Viewer software.

DNA methylation analysis. The DNA methylation status of the D4Z4 region was assayed as previously described[21]. Bisulfite conversion was performed on 1 μg of genomic DNA using the EpiTect Bisulfite Kit (Qiagen) per manufacturer's instructions, and 200 ng of converted genomic DNA was used for PCR. Bisulfite sequencing (BSS) analysis of 52 CpGs in the *DUX4* promoter region of the 4q and 10q D4Z4 repeats was performed using primers BSS167F: TTTTGGGTTGGGTGGAGATTTT and BSS1036R: AACACCRTACCRAACTTACACCCTT, followed by nested PCR with BSS475F: TTAGGAGGGAGGGAGGGAGGTAG and BSS1036R using 10% of the first PCR product. PCR products were cloned into the pGEM-T Easy vector (Promega), sequenced, and analyzed using web-based analysis software BISMA (http://biochem.jacobs-university.de/BDPC/BISMA/)[66] with the default parameters. Standard genomic PCR was performed on non-converted DNA to identify the 4qA, 4qA-L and 4qB chromosome[67]. Specific 4q and 10q haplotypes were identified and assigned as previously described[68,69]. The presence of the DUX4 polyadenylation site was determined by BS-PCR as previously described[42].

Determination of 4q35 and 10q26 D4Z4 array sizes. Peripheral blood leukocytes were embedded in agarose plugs and digested with three different restriction enzymes (EcoRI, EcoRI/BlnI, and XapI). Restriction fragments were separated by pulse field gel electrophoresis (PFGE) and sized and visualized by Southern blot with a p13E-11 probe, and in some subjects, a D4Z4 probe for confirmation[70].

Gene suppression and *in vivo* complementation of zebrafish embryos. Splice blocking morpholinos (MO)s targeting the *Danio rerio smchd1* exon 3 splice donor (e3i3; 5'-AGGTGTGATTTCAGACTTACGCAAC-3') or exon 5 splice donor (e5i5; 5'-TGATTATGAAGACCGCACCTTTGAA-3') were designed and synthesized by Gene Tools LLC (Philomath, Oregon). To determine the optimal MO dose for *in vivo* complementation studies, we injected increasing doses (3 ng, 6 ng, and 9 ng of each MO; 1 nl MO injected per embryo; 1-2 cell stage) into *-1.4col1a1:egfp*[71] embryos harvested from natural mating of heterozygous transgenic adults maintained on an AB background. To determine MO efficiency, we used Trizol (ThermoFisher) to extract total RNA from embryos at 1 day post-fertilization (dpf) according to manufacturer's instructions. Resulting total RNA was reverse transcribed into cDNA using the Superscript III Reverse Transcriptase kit (ThermoFisher), and was used as template in RT-PCR reactions to amplify regions flanking MO target sites. RT-PCR products were gel-purified using the QIAquick gel extraction kit (Qiagen), cloned (TOPO-TA; Invitrogen), and plasmid purified from individual colonies was Sanger sequenced according to standard protocols to identify the precise alteration of endogenous transcript. For rescue experiments, a wild-type (WT) human *SMCHD1* ORF (NM_015295) construct was obtained commercially (OriGene Technologies) and subcloned into the pCS2+ vector. Point mutations were introduced into pCS2+ vectors as described[72] and all vectors were sequence confirmed. WT and variant *SMCHD1* constructs were linearized with *NotI*, and mRNA was transcribed using the mMessage mMachine kit SP6 transcription kit (ThermoFisher). Unless otherwise noted, 9 ng MO (either e3i3 or e5i5) was used in parallel or in combination with 25 pg *SMCHD1* mRNA for *in vivo* complementation studies.

CRISPR/Cas9 genome editing in zebrafish embryos. We used CHOPCHOP (http://chopchop.cbu.uib.no/) to identify a guide (g)RNA targeting sequence a within the smchd1 coding regions (5' GAGATGTCGAAAGTCCGCGG 3'). Guide RNAs were in vitro transcribed using the GeneArt precision gRNA synthesis kit (ThermoFisher) according to manufacturer's instructions. Zebrafish embryos were obtained from -1.4col1a1:egfp embryos harvested from natural mating of heterozygous transgenic adults maintained on an AB background; 1 nl of injection cocktail containing 100 pg/nl gRNA and 200 pg/nl Cas9 protein (PNA Bio) were injected into the cell of embryos at the one-cell stage. To determine targeting efficiency in founder (F0) mutants, we extracted genomic DNA from 2 dpf embryos and PCR-amplified the region flanking the gRNA target site. PCR products were denatured, reannealed slowly and separated on a 15% TBE 1.0 mm precast polyacrylamide gel; it was incubated in ethidium bromide and imaged on a ChemiDoc system (BioRad) to visualize hetero/homoduplexes. To estimate the percent mosaicism of smchd1 F0 mutants (n=5), PCR products were gel purified (Qiagen), and cloned into a TOPO-TA vector (ThermoFisher). Plasmid was prepped from individual colonies (n=10-12 colonies/embryo), and Sanger sequenced according to standard procedures.

Phenotypic analyses in zebrafish. To study craniofacial structures (cartilage or eye development), larval batches were reared at 28^0C and imaged live at 3 dpf using the Vertebrate Automated Screening Technology Bioimager (VAST; software version 1.2.2.8; Union Biometrica) mounted on an AxioScope A1 (Zeiss) microscope using an Axiocam 503 monochromatic camera and Zen Pro 2012 software (Zeiss). Fluorescence imaging of GFP positive cells on ventrally positioned larvae was conducted as described[73]. In parallel, we obtained lateral bright-field images of whole larvae using the VAST onboard camera. To evaluate gonadotropin-releasing hormone (GnRH) neurons, 1.5 dpf embryos were dechorionated and fixed in a solution of 4% paraformaldehyde (PFA) and 7% picric acid for 2 hours at room temperature. Embryos were then washed with a solution of phosphate buffered saline with 0.1% Triton X-100 (PBS-T) and stored at 4^0C until staining. For whole-mount immunostaining, embryos were washed briefly with 0.1% trypsin in PBS; washed in PBS-T; and dehydrated at -20^0C in pre-chilled 100% acetone for 15 min. Next, embryos were washed in PBS-T; blocked in a solution of 2% BSA, 1% DMSO, 0.5% Triton-X100, and 5% calf serum for 1 hour at room temperature. We used rabbit anti-GnRH antibody (1:500 dilution; Sigma) for primary detection. Following overnight incubation of primary antibody, we washed with blocking solution, and incubated with AlexaFluor 555 anti-rabbit secondary antibody (1:500; ThermoFisher) for 2 hours at room temperature. Images were acquired manually with an AxioZoom.V16 microscope and Axiocam 503 monochromatic camera, and were z-stacked using Zen Pro 2012 software (Zeiss). Cartilage structure, eye area, and GnRH neuron projection length was measured using ImageJ software (NIH); pairwise comparisons to determine statistical significance were calculated using a student's t-test. For ceratobranchial pair counts, we used a χ^2 test to determine statistical significance. All experiments were repeated at least twice.

REFERENCES

1. Bosma, J.F., Henkin, R.I., Christiansen, R.L. & Herdt, J.R. Hypoplasia of the nose and eyes, hyposmia, hypogeusia, and hypogonadotrophic hypogonadism in two males. *J Craniofac Genet Dev Biol* **1**, 153-84 (1981).
2. Hogan, B.L. *et al.* Small eyes (Sey): a homozygous lethal mutation on chromosome 2 which affects the differentiation of both lens and nasal placodes in the mouse. *J Embryol Exp Morphol* **97**, 95-110 (1986).
3. Glaser, T. *et al.* PAX6 gene dosage effect in a family with congenital cataracts, aniridia, anophthalmia and central nervous system defects. *Nat Genet* **7**, 463-71 (1994).
4. Schmidt-Sidor, B. *et al.* Malformations of the brain in two fetuses with a compound heterozygosity for two PAX6 mutations. *Folia Neuropathol* **47**, 372-82 (2009).
5. Solomon, B.D. *et al.* Compound heterozygosity for mutations in PAX6 in a patient with complex brain anomaly, neonatal diabetes mellitus, and microophthalmia. *Am J Med Genet A* **149A**, 2543-6 (2009).
6. Cho, C.-H., Shakibaei, M., Merker, H.-J. & Klein, M. The rare malformation of nasal aplasia. *Mund-, Kiefer- und Gesichtschirurgie* **10**, 107-118 (2006).
7. Thiele, H., Musil, A., Nagel, F. & Majewski, F. Familial arhinia, choanal atresia, and microphthalmia. *Am J Med Genet* **63**, 310-3 (1996).
8. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).
9. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv* (2015).
10. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).
11. Chen, K. *et al.* The epigenetic regulator Smchd1 contains a functional GHKL-type ATPase domain. *Biochem J* (2016).
12. Blewitt, M.E. *et al.* SmcHD1, containing a structural-maintenance-of-chromosomes hinge domain, has a critical role in X inactivation. *Nat Genet* **40**, 663-9 (2008).
13. Chen, K. *et al.* Genome-wide binding and mechanistic analyses of Smchd1-mediated epigenetic regulation. *Proc Natl Acad Sci U S A* **112**, E3535-44 (2015).
14. Gendrel, A.V. *et al.* Epigenetic functions of smchd1 repress gene clusters on the inactive X chromosome and on autosomes. *Mol Cell Biol* **33**, 3150-65 (2013).
15. Mould, A.W. *et al.* Smchd1 regulates a subset of autosomal genes subject to monoallelic expression in addition to being critical for X inactivation. *Epigenetics Chromatin* **6**, 19 (2013).
16. Lemmers, R.J. *et al.* Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat Genet* **44**, 1370-4 (2012).
17. Hirano, T. At the heart of the chromosome: SMC proteins in action. *Nat Rev Mol Cell Biol* **7**, 311-22 (2006).
18. Lemmers, R.J. *et al.* Inter-individual differences in CpG methylation at D4Z4 correlate with clinical variability in FSHD1 and FSHD2. *Hum Mol Genet* **24**, 659-69 (2015).
19. van den Boogaard, M.L. *et al.* Double SMCHD1 variants in FSHD2: the synergistic effect of two SMCHD1 variants on D4Z4 hypomethylation and disease penetrance in FSHD2. *Eur J Hum Genet* **24**, 78-85 (2016).

20. Lemmers, R.J. *et al.* Hemizygosity for SMCHD1 in Facioscapulohumeral Muscular Dystrophy Type 2: Consequences for 18p Deletion Syndrome. *Hum Mutat* **36**, 679-83 (2015).
21. Jones, T.I. *et al.* Identifying diagnostic DNA methylation profiles for facioscapulohumeral muscular dystrophy in blood and saliva using bisulfite sequencing. *Clin Epigenetics* **6**, 23 (2014).
22. Jones, T.I. *et al.* Individual epigenetic status of the pathogenic D4Z4 macrosatellite correlates with disease in facioscapulohumeral muscular dystrophy. *Clin Epigenetics* **7**, 37 (2015).
23. Steven, C. *et al.* Molecular characterization of the GnRH system in zebrafish (Danio rerio): cloning of chicken GnRH-II, adult brain expression patterns and pituitary content of salmon GnRH and chicken GnRH-II. *Gen Comp Endocrinol* **133**, 27-37 (2003).
24. Liu, C. *et al.* A secreted WNT-ligand-binding domain of FZD5 generated by a frameshift mutation causes autosomal dominant coloboma. *Hum Mol Genet* **25**, 1382-91 (2016).
25. Chassaing, N. *et al.* Targeted resequencing identifies PTCH1 as a major contributor to ocular developmental anomalies and extends the SOX2 regulatory network. *Genome Res* **26**, 474-85 (2016).
26. Yahyavi, M. *et al.* ALDH1A3 loss of function causes bilateral anophthalmia/microphthalmia and hypoplasia of the optic nerve and optic chiasm. *Hum Mol Genet* **22**, 3250-8 (2013).
27. Whitlock, K.E., Illing, N., Brideau, N.J., Smith, K.M. & Twomey, S. Development of GnRH cells: Setting the stage for puberty. *Mol Cell Endocrinol* **254-255**, 39-50 (2006).
28. Zohar, Y., Munoz-Cueto, J.A., Elizur, A. & Kah, O. Neuroendocrinology of reproduction in teleost fish. *Gen Comp Endocrinol* **165**, 438-55 (2010).
29. Abraham, E., Palevitch, O., Gothilf, Y. & Zohar, Y. The zebrafish as a model system for forebrain GnRH neuronal development. *Gen Comp Endocrinol* **164**, 151-60 (2009).
30. Sharpe, J. *et al.* Optical projection tomography as a tool for 3D microscopy and gene expression studies. *Science* **296**, 541-5 (2002).
31. Blewitt, M.E. *et al.* An N-ethyl-N-nitrosourea screen for genes involved in variegation in the mouse. *Proc Natl Acad Sci U S A* **102**, 7629-34 (2005).
32. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. & Sternberg, M.J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-58 (2015).
33. Chen, J., Bardes, E.E., Aronow, B.J. & Jegga, A.G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* **37**, W305-11 (2009).
34. Hall, J.G. Pena-Shokeir phenotype (fetal akinesia deformation sequence) revisited. *Birth Defects Res A Clin Mol Teratol* **85**, 677-94 (2009).
35. Solomon, B.D., Gropman, A. & Muenke, M. Holoprosencephaly Overview. in *GeneReviews(R)* (eds. Pagon, R.A. *et al.*) (Seattle (WA), 1993).
36. Lederer, D. *et al.* Deletion of KDM6A, a histone demethylase interacting with MLL2, in three patients with Kabuki syndrome. *Am J Hum Genet* **90**, 119-24 (2012).
37. Lindgren, A.M. *et al.* Haploinsufficiency of KDM6A is associated with severe psychomotor retardation, global growth restriction, seizures and cleft palate. *Hum Genet* **132**, 537-52 (2013).
38. de Greef, J.C. *et al.* Clinical features of facioscapulohumeral muscular dystrophy 2. *Neurology* **75**, 1548-54 (2010).

39. van Deutekom, J.C. *et al.* Evidence for subtelomeric exchange of 3.3 kb tandemly repeated units between chromosomes 4q35 and 10q26: implications for genetic counselling and etiology of FSHD1. *Hum Mol Genet* **5**, 1997-2003 (1996).
40. Clapp, J. *et al.* Evolutionary conservation of a coding function for D4Z4, the tandem DNA repeat mutated in facioscapulohumeral muscular dystrophy. *Am J Hum Genet* **81**, 264-79 (2007).
41. Mitsuhashi, H., Mitsuhashi, S., Lynn-Jones, T., Kawahara, G. & Kunkel, L.M. Expression of DUX4 in zebrafish development recapitulates facioscapulohumeral muscular dystrophy. *Hum Mol Genet* **22**, 568-77 (2013).
42. Calandra, P. *et al.* Allele-specific DNA hypomethylation characterises FSHD1 and FSHD2. *J Med Genet* **53**, 348-55 (2016).
43. van den Boogaard, M.L. *et al.* Mutations in DNMT3B Modify Epigenetic Repression of the D4Z4 Repeat and the Penetrance of Facioscapulohumeral Dystrophy. *Am J Hum Genet* **98**, 1020-9 (2016).
44. Weemaes, C.M. *et al.* Heterogeneous clinical presentation in ICF syndrome: correlation with underlying gene defects. *Eur J Hum Genet* **21**, 1219-25 (2013).
45. Fokkema, I.F. *et al.* LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat* **32**, 557-63 (2011).
46. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401-4 (2012).
47. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013).
48. Petersen, B., Petersen, T.N., Andersen, P., Nielsen, M. & Lundegaard, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol* **9**, 51 (2009).
49. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat Methods* **12**, 7-8 (2015).
50. Heffernan, R. *et al.* Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep* **5**, 11476 (2015).
51. Miller, D.T. *et al.* Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* **86**, 749-64 (2010).
52. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).
53. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 1-33 (2013).
54. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-73 (2010).
55. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
56. Turner, S.D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv* (2014).
57. Blumenthal, I. *et al.* Transcriptional consequences of 16p11.2 deletion and duplication in mouse cortex and multiplex autism families. *Am J Hum Genet* **94**, 870-83 (2014).

58. Sugathan, A. *et al.* CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc Natl Acad Sci U S A* **111**, E4468-77 (2014).
59. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873-81 (2010).
60. DeLuca, D.S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530-2 (2012).
61. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184-5 (2012).
62. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
63. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
64. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *2015* **67**, 48 (2015).
65. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-9 (2006).
66. Rohde, C., Zhang, Y., Reinhardt, R. & Jeltsch, A. BISMA--fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences. *BMC Bioinformatics* **11**, 230 (2010).
67. Lemmers, R.J. *et al.* A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* **329**, 1650-3 (2010).
68. Lemmers, R.J. *et al.* Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. *Am J Hum Genet* **81**, 884-94 (2007).
69. Lemmers, R.J. *et al.* Worldwide population analysis of the 4q and 10q subtelomeres identifies only four discrete interchromosomal sequence transfers in human evolution. *Am J Hum Genet* **86**, 364-77 (2010).
70. Lemmers, R.J., O'Shea, S., Padberg, G.W., Lunt, P.W. & van der Maarel, S.M. Best practice guidelines on genetic diagnostics of Facioscapulohumeral muscular dystrophy: workshop 9th June 2010, LUMC, Leiden, The Netherlands. *Neuromuscul Disord* **22**, 463-70 (2012).
71. Kague, E. *et al.* Skeletogenic fate of zebrafish cranial and trunk neural crest. *PLoS One* **7**, e47394 (2012).
72. Niederriter, A.R. *et al.* In vivo modeling of the morbid human genome using Danio rerio. *J Vis Exp*, e50338 (2013).
73. Isrie, M. *et al.* Mutations in Either TUBB or MAPRE2 Cause Circumferential Skin Creases Kunze Type. *Am J Hum Genet* **97**, 790-800 (2015).